

CLAIMS

1. A method for clustering queries, the method comprising:
identifying a same document and/or a plurality of similar documents
selected by a user in response to a plurality of queries; and
responsive to identifying the same document and/or the similar documents,
generating a query cluster to indicate that the queries are similar independent of
whether individual ones of the queries comprise similar composition with respect
to other ones of the queries.

2. A method as recited in claim 1, wherein the queries comprise a well
formed natural language question, a keyword, or a phrase.

3. A method as recited in claim 1, wherein the query cluster is used to
disambiguate a word or phrase in a query of the queries.

4. A method as recited in claim 1, further comprising determining that
the queries are similar based on similar keyword or phrase composition.

1 5. A method as recited in claim 1, wherein identifying the same
2 document and/or the similar documents further comprises:

3 determining the similar documents by evaluating a set of selected similar
4 documents chosen responsive to queries p and q of the queries, wherein
5 documents $D_C(.)$ is a subset of a result list $D(.)$ according to the following:

$$6 \quad D_C(p) = \{ d_{p1}, d_{p2}, \dots, d_{pi} \} \subseteq D(p)$$

$$7 \quad D_C(q) = \{ d_{q1}, d_{q2}, \dots, d_{qj} \} \subseteq D(q);$$

8 wherein similarity based on selection of documents is based on:

9 If $D_C(p) \cap D_C(q) = \{ d_{pq1}, d_{pq2}, \dots, d_{pqk} \} \neq \emptyset$, then documents d_{pq1} ,
10 d_{pq2}, \dots, d_{pqk} represent a set of common topics of queries p and q , and,

11 whereby the similar documents between queries p and q is determined by
12 $D_C(p) \cap D_C(q)$.

13
14 6. A method as recited in claim 1, further comprising constructing a
15 thesaurus comprising a plurality of synsets, wherein each synset comprises one or
16 more query clusters.

17
18 7. A method as recited in claim 1, wherein identifying the same
19 document and/or the similar documents further comprises determining the similar
20 documents based on a proportionality of commonly selected individual
21 documents.

1 8. A method as recited in claim 7, wherein identifying the same
2 document and/or the similar documents further comprises:

3 determining the similar documents based on a proportionality of commonly
4 selected individual documents, such that:

$$5 \quad \text{similarity}_{\text{single_doc}}(p, q) = \frac{RD(p, q)}{\text{Max}(rd(p), rd(q))},$$

6 wherein $rd(.)$ is the number of clicked documents for a query of the queries,
7 and wherein $RD(p, q)$ is the number of document selections in common.

8
9 9. A method as recited in claim 1, wherein identifying the same
10 document and/or the similar documents further comprises:

11 determining the similar documents based on a hierarchical positioning
12 between individual ones of a plurality of documents commonly selected across the
13 queries.
14
15
16
17
18
19
20
21
22
23
24
25

10. A method as recited in claim 9:

wherein $F(d_i, d_j)$ is a lowest common parent node for documents d_i and d_j ;

wherein $L(x)$ is a level of a node x ;

wherein L_Total identifies a total number of levels in a hierarchy; and

wherein a similarity between two documents is defined as follows:

$$s(d_i, d_j) = \frac{L(F(d_i, d_j)) - 1}{L_Total - 1}, \text{ such that}$$

$$s(d_i, d_i) = 1; \text{ and } s(d_i, d_j) = 0 \text{ if } F(d_i, d_j) = \text{root}; \text{ and}$$

the method further comprises:

incorporating $s(d_i, d_j)$ into a calculation of query similarity, wherein.

d_i ($1 \leq i \leq m$) and d_j ($1 \leq j \leq n$) be a set of selected documents for queries p and q respectively such that:

$$similarity_{hierarchy}(p, q) = \frac{1}{2} \times \left(\frac{\sum_{i=1}^m (\max_{j=1}^n s(d_i, d_j))}{rd(p)} + \frac{\sum_{j=1}^n (\max_{i=1}^m s(d_i, d_j))}{rd(q)} \right)$$

11. Computer-readable media comprising computer-executable instructions for identifying similar queries, the computer-executable instructions comprising instructions for:

identifying a same document and/or a plurality of similar documents selected by a user in response to a plurality of queries; and

responsive to identifying the same document and/or the similar documents, generating a query cluster to indicate that the queries are similar independent of whether individual ones of the queries comprise similar composition with respect to other ones of the queries.

1
2 **12.** Computer-readable media as recited in claim 11, wherein the queries
3 comprise a well formed natural language question, a keyword, or a phrase.
4

5 **13.** Computer-readable media as recited in claim 11, wherein the query
6 cluster is used to disambiguate a word or phrase in a query of the queries.
7

8 **14.** Computer-readable media as recited in claim 11, wherein the
9 computer-executable instructions further comprise instructions for determining
10 that the queries are similar based on similar keyword or phrase composition.
11

12 **15.** Computer-readable media as recited in claim 11, wherein the
13 instructions for identifying the same document and/or the similar documents
14 further comprise instructions for:

15 determining the similar documents by evaluating a set of selected similar
16 documents chosen responsive to queries p and q of the queries, wherein
17 documents $D_C(.)$ is a subset of a result list $D(.)$ according to the following:

18
$$D_C(p) = \{ d_{p1}, d_{p2}, \dots, d_{pi} \} \subseteq D(p)$$

19
$$D_C(q) = \{ d_{q1}, d_{q2}, \dots, d_{qj} \} \subseteq D(q);$$

20 wherein similarity based on selection of documents is based on:

21 If $D_C(p) \cap D_C(q) = \{ d_{pq1}, d_{pq2}, \dots, d_{pqk} \} \neq \emptyset$, then documents $d_{pq1},$
22 d_{pq2}, \dots, d_{pqk} represent a set of common topics of queries p and q , and,

23 whereby the similar documents between queries p and q is determined by
24 $D_C(p) \cap D_C(q).$
25

1 **19.** Computer-readable media as recited in claim 11, wherein the
2 instructions for identifying the same document and/or the similar documents
3 further comprise instructions for:

4 determining the similar documents based on a hierarchical positioning
5 between individual ones of a plurality of documents commonly selected across the
6 queries.

7
8 **20.** Computer-readable media as recited in claim 19:

9 wherein $F(d_i, d_j)$ is a lowest common parent node for documents d_i and d_j ;

10 wherein $L(x)$ is a level of a node x ;

11 wherein L_Total identifies a total number of levels in a hierarchy; and

12 wherein a similarity between two documents is defined as follows:

13
$$s(d_i, d_j) = \frac{L(F(d_i, d_j)) - 1}{L_Total - 1}, \text{ such that}$$

14
$$s(d_i, d_j) = 1; \text{ and } s(d_i, d_j) = 0 \text{ if } F(d_i, d_j) = \text{root}; \text{ and}$$

15 wherein the computer-executable instructions further comprise instructions
16 for:

17 incorporating $s(d_i, d_j)$ into a calculation of query similarity, wherein.
18 d_i ($1 \leq i \leq m$) and d_j ($1 \leq j \leq n$) be a set of selected documents for queries p and q
19 respectively such that:

20
21
$$similarity_{hierarchy}(p, q) = \frac{1}{2} \times \left(\frac{\sum_{i=1}^m (\max_{j=1}^n s(d_i, d_j))}{rd(p)} + \frac{\sum_{j=1}^n (\max_{i=1}^m s(d_i, d_j))}{rd(q)} \right)$$

22
23
24
25

1 **21.** A computing device comprising:

2 a processor coupled to a memory, the memory comprising computer
3 executable instructions, the processor being configured to fetch and execute the
4 computer-executable instructions for:

5 identifying a same document and/or a plurality of similar documents
6 selected by a user in response to a plurality of queries; and

7 responsive to identifying the same document and/or the similar
8 documents, generating a query cluster to indicate that the queries are similar
9 independent of whether individual ones of the queries comprise similar
10 composition with respect to other ones of the queries.

11
12 **22.** A computing device as recited in claim 21, wherein the queries
13 comprise a well formed natural language question, a keyword, or a phrase.

14
15 **23.** A computing device as recited in claim 21, wherein the query cluster
16 is used to disambiguate a word or phrase in a query of the queries.

17
18 **24.** A computing device as recited in claim 21, wherein the computer-
19 executable instructions further comprise instructions for determining that the
20 queries are similar based on similar keyword or phrase composition.
21
22
23
24
25

1 **25.** A computing device as recited in claim 21, wherein the instructions
2 for identifying the same document and/or the similar documents further comprise
3 instructions for:

4 determining the similar documents by evaluating a set of selected similar
5 documents chosen responsive to queries p and q of the queries, wherein
6 documents $D_C(.)$ is a subset of a result list $D(.)$ according to the following:

$$D_C(p) = \{ d_{p1}, d_{p2}, \dots, d_{pi} \} \subseteq D(p)$$

$$D_C(q) = \{ d_{q1}, d_{q2}, \dots, d_{qj} \} \subseteq D(q);$$

7 wherein similarity based on selection of documents is based on:

8 If $D_C(p) \cap D_C(q) = \{ d_{pq1}, d_{pq2}, \dots, d_{pqk} \} \neq \emptyset$, then documents $d_{pq1},$
9 d_{pq2}, \dots, d_{pqk} represent a set of common topics of queries p and q , and,

10 whereby the similar documents between queries p and q is determined by
11 $D_C(p) \cap D_C(q)$.

12
13
14
15 **26.** A computing device as recited in claim 21, wherein the computer-
16 executable instructions further comprise instructions for constructing a thesaurus
17 comprising a plurality of synsets, wherein each synset comprises one or more
18 query clusters.

19
20 **27.** A computing device as recited in claim 21, wherein the instructions
21 for identifying the same document and/or the similar documents further comprise
22 instructions for determining the similar documents based on a proportionality of
23 commonly selected individual documents.

30. A computing device as recited in claim 29:

wherein $F(d_i, d_j)$ is a lowest common parent node for documents d_i and d_j ;

wherein $L(x)$ is a level of a node x ;

wherein L_Total identifies a total number of levels in a hierarchy; and

wherein a similarity between two documents is defined as follows:

$$s(d_i, d_j) = \frac{L(F(d_i, d_j)) - 1}{L_Total - 1}, \text{ such that}$$

$$s(d_i, d_i) = 1; \text{ and } s(d_i, d_j) = 0 \text{ if } F(d_i, d_j) = \text{root}; \text{ and}$$

wherein the computer-executable instructions further comprise instructions for:

incorporating $s(d_i, d_j)$ into a calculation of query similarity, wherein.

d_i ($1 \leq i \leq m$) and d_j ($1 \leq j \leq n$) be a set of selected documents for queries p and q respectively such that:

$$similarity_{hierarchy}(p, q) = \frac{1}{2} \times \left(\frac{\sum_{i=1}^m (\max_{j=1}^n s(d_i, d_j))}{rd(p)} + \frac{\sum_{j=1}^n (\max_{i=1}^m s(d_i, d_j))}{rd(q)} \right)$$